Lecture 6: Returns and basic statistics

Ziwen Ye

Stevens Institute of Technology

zye2@stevens.edu

October 7, 2018

1. Return
    - Definition
2. Moments
    - Mean and Variance
    - Skewness
    - Kurtosis
3. Statistic related test
    - Jarque-Bera Test
    - Covariance and Correlation
4. Time-series related test
    - Stationary
    - ADF test
    - Autocorrelation

Return

Simple Return

$$1 + R_t = \frac{P_t}{P_{t-1}}$$

$$\therefore R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Log Return (continuously compounded return)

$$r_t = log(1 + R_t) = log(P_t / P_{t-1})$$

$$= log(P_t) - log(P_{t-1})$$

Return

For multiple time intervals, from t-k to t

Simple

$$1 + R_{t[k]} = \frac{P_t}{P_{t-k}} = \frac{P_t}{P_{t-1}} \frac{P_{t-1}}{P_{t-2}} \cdots \frac{P_{t-k+1}}{P_{t-k}}$$

$$= (1 + R_t)(1 + R_{t-1}) ... (1 + R_{t-k+1}) = \prod_{j=0}^{k-1} (1 + R_{t-j})$$

Log

$$r_{t[k]} = log(1 + R_{t[k]}) = log[\frac{P_t}{P_{t-1}} \frac{P_{t-1}}{P_{t-2}} \cdots \frac{P_{t-k+1}}{P_{t-k}}]$$

$$= log(\frac{P_t}{P_{t-1}}) + \cdots + log(\frac{P_{t-k+1}}{P_{t-k}}) = \sum_{j=1}^{k-1} r_j$$

Return

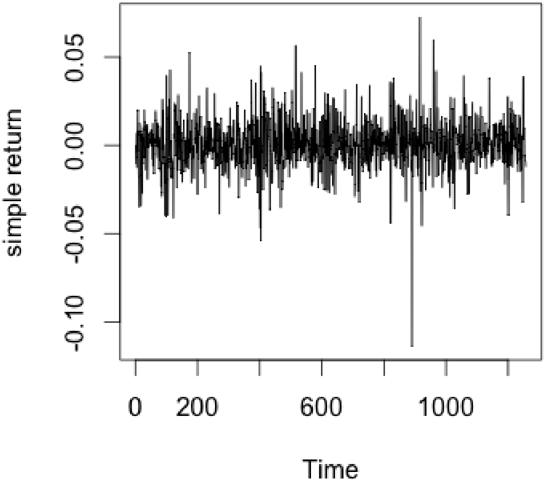
Calculate return in R.

Example

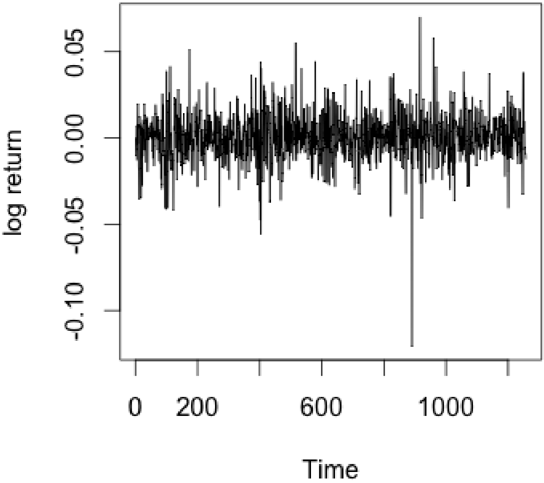
```
> msft <- read.csv("msft.csv")
> # reverse the row order
> msft <- msft[nrow(msft):1, ]

> msft.price <- msft$Adj.Close
> msft.pt <- msft.price[2:length(msft.price)]
> msft.pt_1 <- msft.price[1: (length(msft.price)-1) ]
> # calculate simple return and log return
> msft.simple.return <- (msft.pt - msft.pt_1) / msft.pt_1
> msft.log.return <- log(msft.pt) - log(msft.pt_1)
>
> plot.ts(msft.simple.return)
> plot.ts(msft.log.return)
```

Return



Return



Return

Why use return series instead of price series.

- Return is a complete and scale-free summary of an asset.
- Return has more attractive statistical properties. (weakly stationary)

Why we need to assume data is stationary?

Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary (i.e., ”stationarized”) through the use of mathematical transformations. A stationarized series is relatively easy to predict: you simply predict that its statistical properties will be the same in the future as they have been in the past!

Mean and Variance

Assume we have a vector $X = \{x_1, x_2, \dots, x_n\}$

Mean (first moment)

$$E[X] = \frac{x_1 + \cdots + x_n}{n} = \mu$$

Variance (second moment)

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + \mu^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

Sample and population

Variance (Apply to the population)

$$Var[x] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

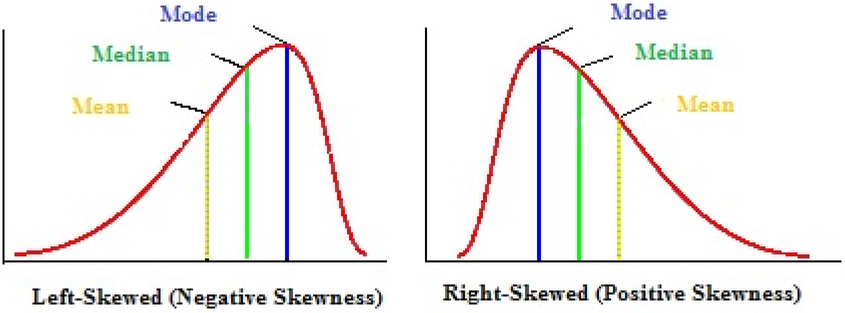
Sample Variance (Apply to the sample)

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

In R, when you use sd() or var() to calculate standard deviation or variance, it always refer to sample standard deviation or sample variance.

Skewness (third moment)

In probability theory and statistics, *skewness* is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Skewness

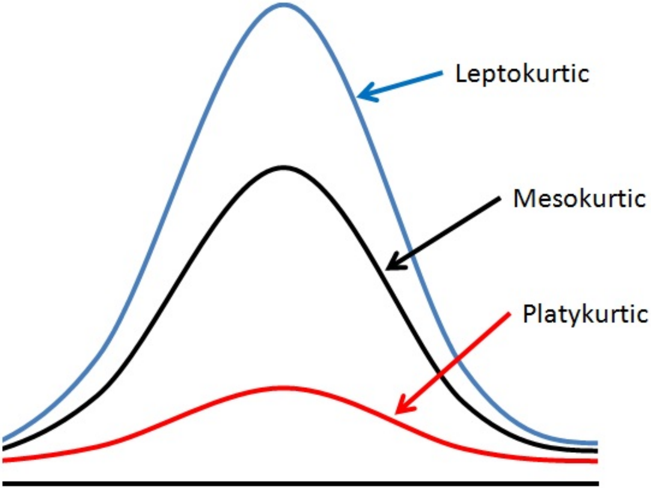
Skewness Equation

$$\gamma = E[(\frac{X-\mu}{\sigma})^3]$$

- For a normal distribution, its skewness is 0
- When skewness is negative, it means the distribution is left skewed
- When skewness is positive, it means the distribution is right skewed

Kurtosis (fourth moment)

In probability theory and statistics, kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable.



Kurtosis

Kurtosis

$$Kurt[X] = E[(\frac{X-\mu}{\sigma})^4]$$

- For normal distribution, its kurtosis is 3
- Excess kurtosis= Kurt[X]-3 (In R you can calculate excess kurtosis)
- If Kurt[X] $>$ 3, it means this distribution has heavy tail(Leptokurtic behavior)
- If Kurt[X] $<$ 3, it means this distribution has short tail(Platykurtic behavior)

Jarque-Bera Test

Usually we assume return would apply to Normal distribution. Jarque-Bera test is a method to test whether sample data have the skewness and kurtosis matching a normal distribution

Null hypothesis

The skewness being zero and the excess kurtosis being zero

Alternative hypothesis

At least one condition is not satisfied

When P-value is larger than 0.05, we fail to reject null hypothesis. When P-value is less than 0.05, we reject our null hypothesis, which imply our dataset is not Normal

Review

Covariance is a measure of how much two random variable change together.

Covariance

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - E[X]E[Y]$$

Covariance

Thus we have: $E[XY] = E[X]E[Y] + Cov(X, Y)$

Correlation Coefficient

$$\rho = Corr(X, Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Stationary

Strictly(or Strongly) Stationary

A time series $\{r_t\}$ is said to be strictly stationary if the joint distribution of $(r_{t_1}, \ldots, r_{t_k})$ is identical to $(r_{t_{1+t}}, \ldots, r_{t_{k+t}})$ for all t, where k is an arbitrary positive integer and $(t_1, \ldots, t_k)$ is a collection of k positive integers

Weakly Stationary

A time series $\{r_t\}$ is said to be weakly stationary if both the mean of r_t and the covariance between r_t and r_{t-l} are time invariant, where l is an arbitrary integer

ADF test

Augmented Dickey-Fuller(ADF) unit-root test can be used to test whether your time series data set is stationary or not.

Null hypothesis

$H_0 =$ At least one unit root is present in a time series sample

Alternative hypothesis

$H_a =$ There is no unit root(This implies dataset is stationary)

Autocorrelation

What is the relationship between the return of today and of yesterday?

The correlation coefficient of r_t and r_{t-l} is called *lag-l* autocorrelation.

Autocorrelation

$$\text{lag 1: } \rho_1 = Corr(r_t, r_{t-1}) = \frac{Cov(r_t,r_{t-1})}{\sigma_t\sigma_{t-1}} = \frac{Cov(r_t,r_{t-1})}{Var(r_t)}$$

$$\text{lag 2: } \rho_2 = Corr(r_t, r_{t-2}) = \frac{Cov(r_t,r_{t-2})}{\sigma_t\sigma_{t-2}} = \frac{Cov(r_t,r_{t-2})}{Var(r_t)}$$

$$\text{lag l: } \rho_l = Corr(r_t, r_{t-l}) = \frac{Cov(r_t,r_{t-l})}{\sigma_t\sigma_{t-l}} = \frac{Cov(r_t,r_{t-l})}{Var(r_t)}$$

We use γ_l represent lag-l covariance of r_t, apparently:

$$\rho_l = \gamma_l/\gamma_0$$